Analysis of genomic and ploidy alterations
Network Reconstruction and Analysis TP3

Mathieu Pont and Lucas Rodrigues Pereira

December 2019

Abstract

1 Introduction

In the context of the Master's Degree on Machine Learning for Data Science, at the University of Paris (Descartes), we have been given the task to discuss some network reconstruction methods, as part of the "Network Reconstruction and Analysis" course. The main idea of network reconstruction is to find causality between variables through a graph.

In order to accomplish the task, we want to analyze genomic alterations on breast tumor from the online Catalog of Somatic Mutations in Cancer (COSMIC). Three different methods have been used: Hill-climbing, PC and MIIC.

2 Experiments and Results

2.1 Preliminary study of the data set

The data set contains 807 samples representing cells (diploid and tetraploid tumoral cells) described by 176 variables, each being a gene (mutated or over/under expressed) except one containing Ploidy information.

8 samples were removed because they have NA values for the Ploidy variable. Moreover, 14 variables were also removed because they had the same value across all samples ('esm1', 'ebf4', 'qscn6l1', 'cenpa', 'kntc2', 'orc6l', 'aytl2', 'peci', 'gstm3', 'cdkn2a', 'cdkn1a', 'foxo1', 'ppp2r2a', 'spdye7p').

After pre-processing the dataset contains 799 samples with 162 variables.

2.2 HC

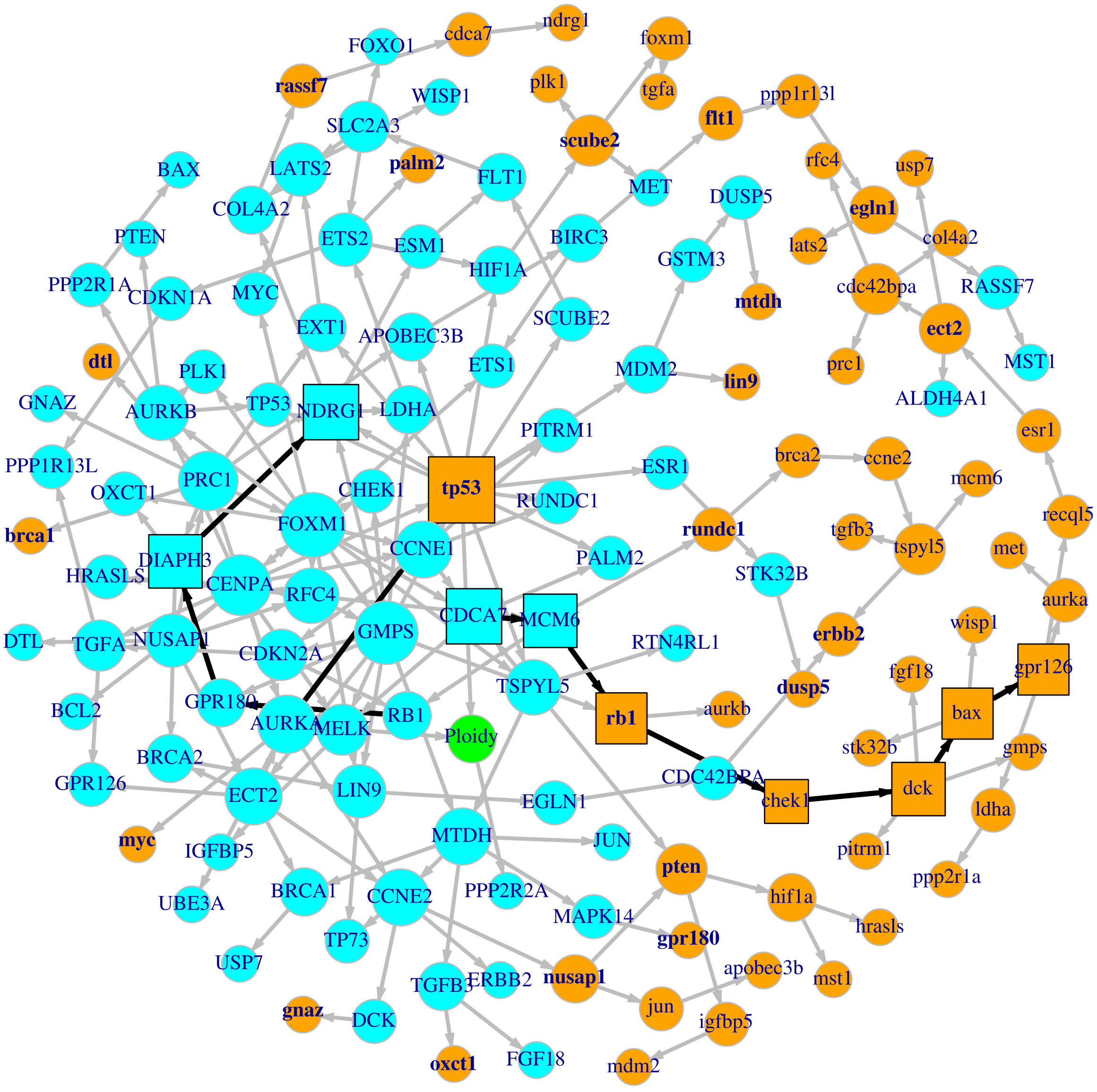


Figure 1: Reconstructed network using the HC algorithm.

The 10 top edges and nodes according to betweenness centrality [1] are highlighted, the edges in black, the nodes with a square shape. Nodes with their name in bold are the mutated genes related the over/under expressed genes. Finally, the size of a node is proportional to its degree and nodes with 0 degree are removed.

The graph given by HC is very dense but the different genes are well grouped together. The density of the graph make it complicated to interpret, moreover we don't have score for each edge to select them. We can however see that ***tp53*** has an important role since it is very related to over/under expressed genes and to ***Ploidy*** and it is the first node in term of betweenness centrality.

A lot of mutated genes have a relation with an over/under expressed genes, since we don't have a score we can't select only the significant ones. However, we can look at those with highest degree, here, ***tp53***, ***rb1*** and ***scube2***.

Three genes are related to ***Ploidy***: ***tp53***, ***AURKA***, ***PPP2R2A***.

tp53	***FOXM1***	***GMPS***	***CENPA***	***PRC1***

Table 1: Hubs of the HC graph (ordered by degree).

tp53	***MCM6***	***dck***	***rb1***	***chek1***
bax	***CDCA7***	***NDRG1***	***DIAPH3***	***gpr126***

Table 2: Top 10 nodes of the HC graph according to betweenness centrality (row-ordered by score).

chek1 → ***dck***	***rb1*** → ***chek1***	***CDCA7*** → ***MCM6***
dck → ***bax***	***bax*** → ***gpr126***	***GPR180*** → ***DIAPH3***
AURKA → ***tp53***	***MCM6*** → ***rb1***	***RB1*** → ***GPR180***
DIAPH3 → ***NDRG1***		

Table 3: Top 10 edges of the HC graph according to betweenness centrality (row-ordered by score).

Among the mutated genes of the top 10 nodes, it is interesting to see that ***rb1***, ***chek1***, ***dck***, ***bax*** and ***gpr126*** form kind of a chain that link the concentrated group of over/under expressed genes with the one of the mutated genes.

Table 3 shows us that most of the top edges are between genes of the same group. Except for two edges, involving ***tp53*** and ***rb1*** meaning that these genes can be seen as a "bridge" between mutated genes and over/under expressed genes.

The betweenness centrality expresses the importance of the position of these nodes in the graph, showing us a link between these mutated and over/under expressed genes. Overall, we saw that ***tp53*** has an important role.

[1] According to [Sunil Kumar Raghavan Unnithan and Jathavedan, 2014], betweenness centrality is a measure of the influence of a vertex over the flow of information between every pair of vertices under the assumption that information primarily flows over the shortest paths between them. In this paper we present betweenness centrality of some important classes of graphs.

2.3 PC

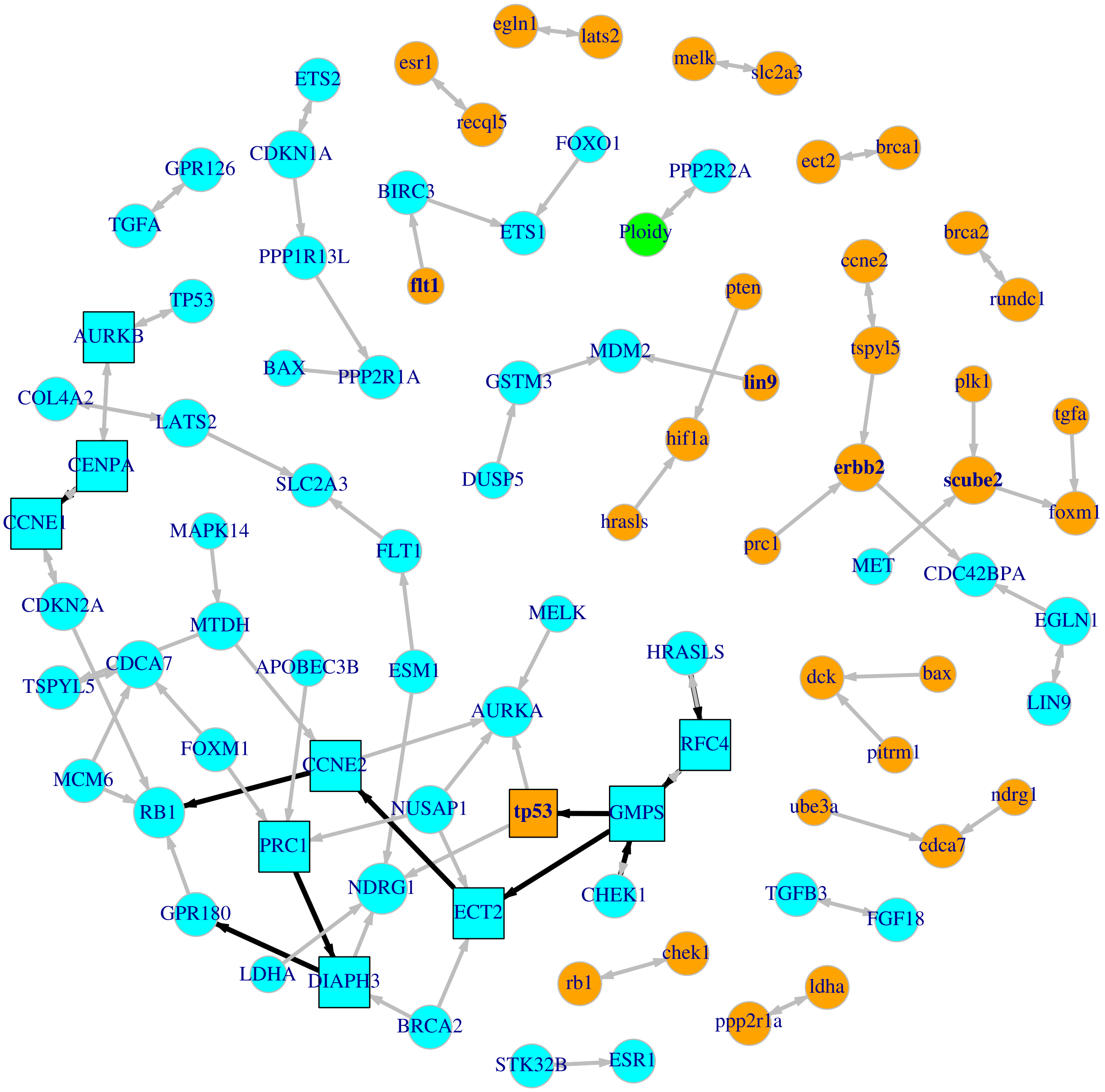


Figure 2: Reconstructed network using the PC algorithm with $\alpha = 0.05$.

The PC graph is less dense than the one of HC, it has many small connected components and a big one. As seen with HC, ***tp53*** is more in relation with over/under expressed genes than the other mutated genes.

According to PC, only one gene is related to ***Ploidy***: ***PPP2R2A***. The edge has a pMax value of 0.0456, it's very near the selected threshold of 0.05, making its significance not very high.

Five mutated genes are related to over/under expressed genes: ***tp53***, ***erbb2***, ***scube2***, ***flt1*** and ***lin9***.

Edge	pMax
GMPS → ***tp53***	0.0007
tp53 → ***NDRG1***	0.0056
tp53 → ***AURKA***	0.0160
erbb2 → ***CDC42BPA***	0.0173
MET → ***scube2***	0.0213
flt1 → ***BIRC3***	0.0314
lin9 → ***MDM2***	0.0398

Table 4: pMax value for each edge between genes of different groups.

The pMax value of edges involving ***tp53*** are very low making their significance more important than those of the bottom of the table. It shows us the importance of relation between ***tp53*** and over/under expressed genes from the statistical test point of view. Information that HC does not gives.

GMPS	***CCNE2***	***ECT2***	***CENPA***	***PRC1***

Table 5: Hubs of the PC graph (ordered by degree).

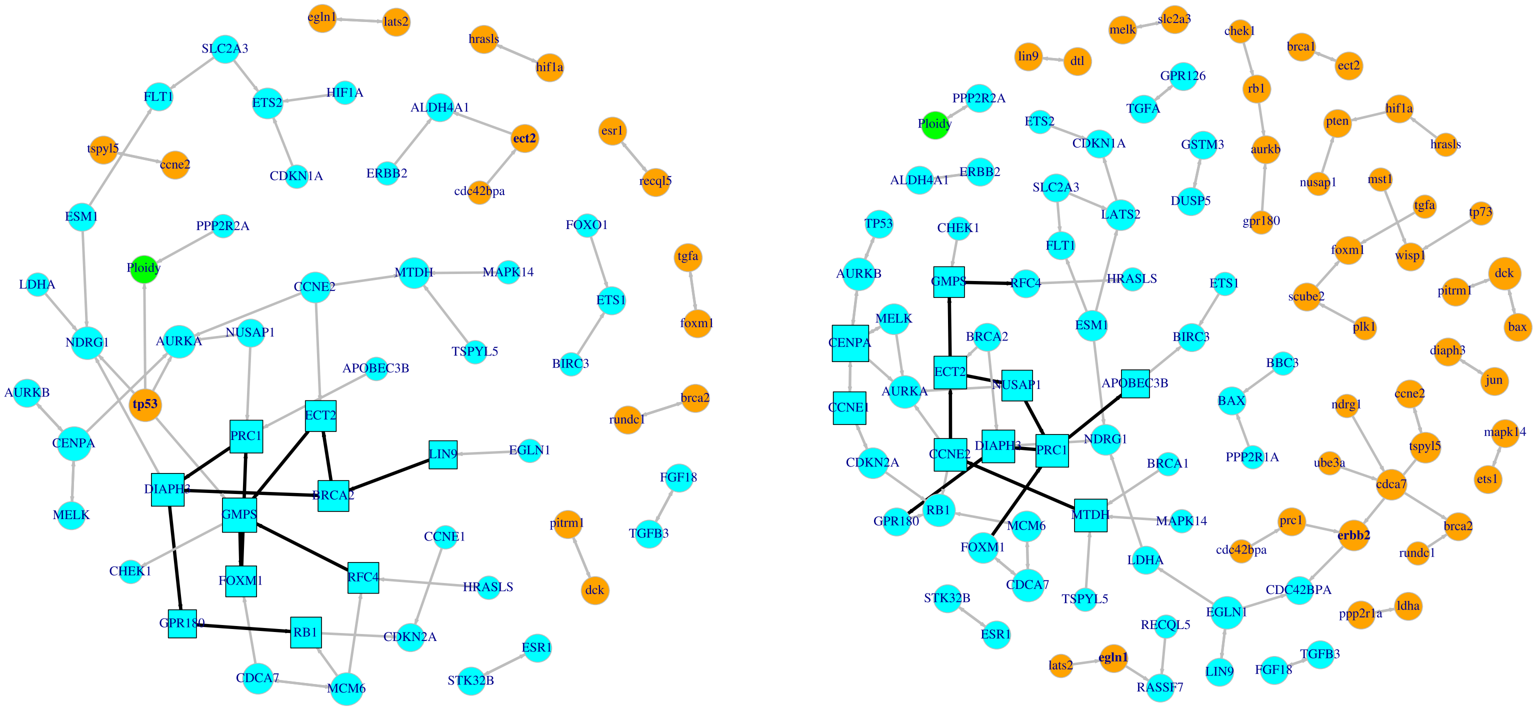
GMPS	***DIAPH3***	***CCNE2***	***ECT2***	***PRC1***
CENPA	***RFC4***	***CCNE1***	***tp53***	***AURKB***

Table 6: Top 10 nodes of the PC graph according to betweenness centrality (row-ordered by score).

GMPS → ***RFC4***	***CCNE2*** → ***ECT2***	***DIAPH3*** → ***PRC1***
ECT2 → ***GMPS***	***tp53*** → ***GMPS***	***GPR180*** → ***DIAPH3***
RB1 → ***CCNE2***	***CCNE1*** → ***CENPA***	***GMPS*** → ***CHEK1***
RFC4 → ***HRASLS***		

Table 7: Top 10 edges of the PC graph according to betweenness centrality (row-ordered by score).

All top edges according to betweenness centrality seems to be between over/under expressed genes. Indeed, we can see that these genes tend to be more connected than the mutated ones, putting them in small connected components. Only one edge connect nodes of both groups, between ***GMPS*** and ***tp53*** showing once again the importance of the latter.



(a) $\alpha = 0.01$

(b) $\alpha = 0.10$

Figure 3: Reconstructed networks using the PC algorithm with different values for α.

In Figure 3a we have run the PC algorithm with a lower threshold, making it more "severe" for choosing edges. It's why we see many small connected components compared to the graph of Figure 3b made with a higher threshold. However, the latter has lost important information about ***tp53*** and its link between ***Ploidy*** and over/under expressed genes.

This threshold is an interesting parameter that allows us to control, in some way, the density of the graph. However, it can lose information as we seen for the graph of Figure 3b with ***tp53***.

2.4 MIIC

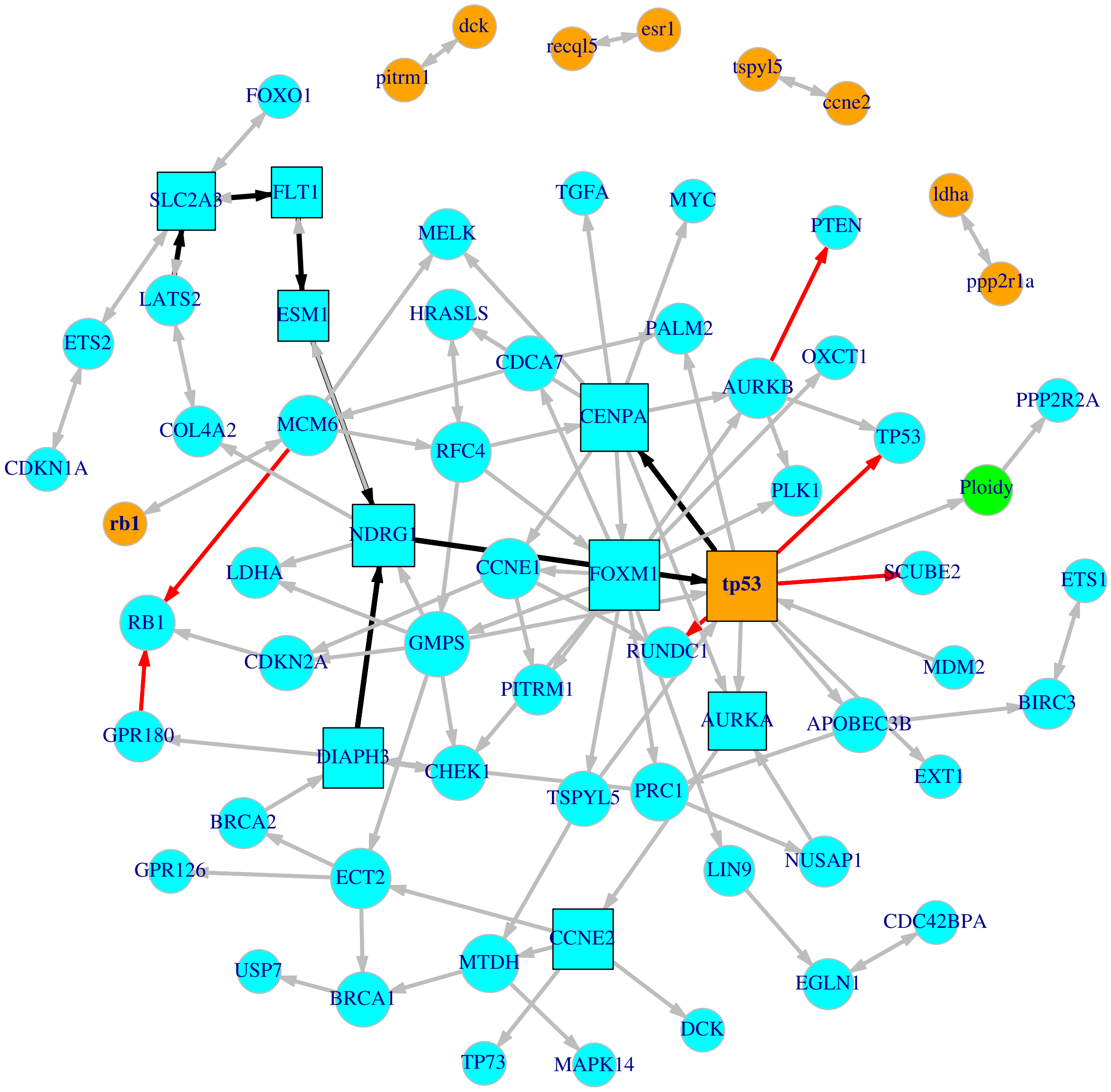


Figure 4: Reconstructed network using the MIIC algorithm with $confidenceShuffle = 100$ and $confidenceThreshold = 0.001$.

The MIIC algorithm, introduced by [Verny et al., 2017], is mainly configured by two parameters (in the *R* package):

- **confidenceShuffle**: According to the MIIC documentation, it is the number of shufflings of the original dataset in order to evaluate the edge specific confidence ratio of all inferred edges. This enforces stability to the reconstructed graph but execution time is longer.
- **confidenceThreshold**: A threshold of confidence to determine if one edge should be kept or discarded.

The MIIC graph is lense dense than with HC, making it easier to interpret. The edges representing repressions (negative partial correlation) are colored in red. Here, only two mutated genes are linked with genes of the other category: ***tp53*** and ***rb1***.

Edge	Log confidence	Partial correlation
tp53 → ***GMPS***	74.20	0.44
tp53 → ***NDRG1***	45.01	0.35
tp53 → ***PALM2***	32.34	0.30
rb1 → ***MCM6***	7.67	0.20

Table 8: Mutated genes related to over/under expressed genes.

Several edges involving ***tp53*** were ommitted due to their total number of 12, only the most significant are present in Table 8. The edge involving ***rb1*** has the lowest log confidence among edges between genes of both groups. Moreover, ***tp53*** has negative partial correlation with ***TP53***, ***MDM2***, ***SCUBE2*** and ***RUNDC1***.

PC also founds relation between ***tp53*** and ***GMPS*** but the edge was in the other direction and both found the same edge between ***tp53*** and ***NDRG1***.

Edge	Log confidence	Partial correlation
tp53 → ***Ploidy***	22.15	0.25
Ploidy → ***PPP2R2A***	8.36	-0.08

Table 9: Variables related to ***Ploidy***.

MIIC found more or less the same relations involving ***Ploidy*** than HC and PC but it also gives us confidence value and partial correlation for each edge. We see that for ***PPP2R2A*** the log confidence is not very high, and the partial correlation near 0, that corroborates what we saw with PC for which the pMax value of this edge was not very low (having therefore a not high significance).

tp53	***FOXM1***	***CENPA***	***GMPS***	***NDRG1***

Table 10: Hubs of the MIIC graph (ordered by degree).

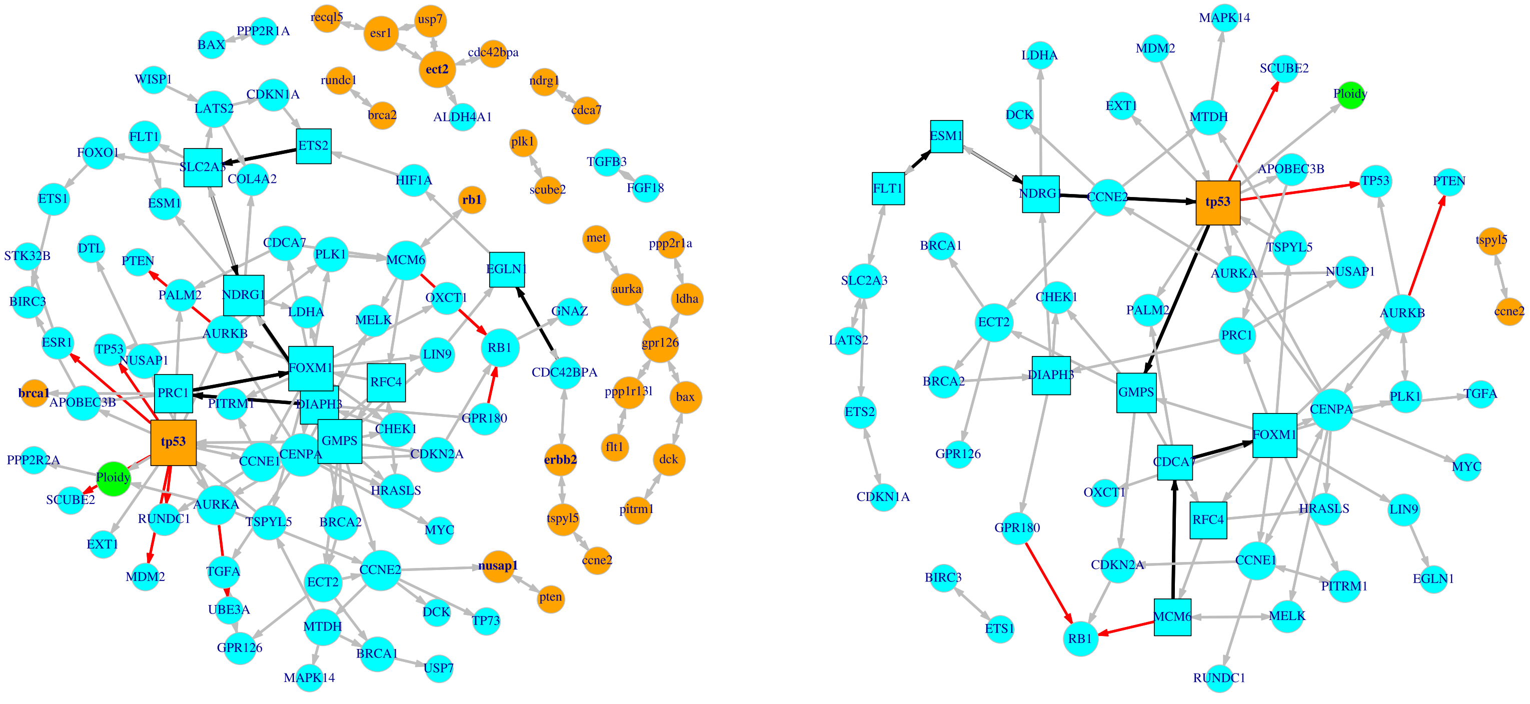
tp53	***NDRG1***	***CENPA***	***DIAPH3***	***ESM1***
FOXM1	***SLC2A3***	***FLT1***	***CCNE2***	***AURKA***

Table 11: Top 10 nodes of the MIIC graph according to betweenness centrality (row-ordered by score).

ESM1 → ***NDRG1***	***NDRG1*** → ***tp53***	***FLT1*** → ***ESM1***
SLC2A3 → ***FLT1***	***tp53*** → ***CENPA***	***tp53*** → ***CENPA***
DIAPH3 → ***NDRG1***	***DIAPH3*** → ***NDRG1***	***LATS2*** → ***SLC2A3***
AURKA → ***CCNE2***		

Table 12: Top 10 edges of the MIIC graph according to betweenness centrality (row-ordered by score).

Like for PC, most of the top edges are between over/under expressed genes and the only mutated genes appearing in the list is ***tp53***.



(a) $confidenceThreshold = 0.01$

(b) $confidenceThreshold = 0.0001$

Figure 5: Reconstructed networks using the MIIC algorithm with different values for $confidenceThreshold$ (with $confidenceShuffle = 100$).

As seen in Figure 5 less edges are selected with a lower confidence threshold (the algorithm is more "severe" in its selection). It is why we can see few edges in the graph of Figure 5b, moreover many mutated genes became unlinked making them not appearing in the final graph.

The $confidenceShuffle$ parameter does not have an impact on the graph but more on the confidence value for each edge.

3 Conclusion

The first thing that caught our attention was the different results that each approach yielded, despite using the same dataset. Even if they all have found that ***tp53*** has an important role, each graph has its particularity. Moreover, not being an expert in the subject makes it harder to evaluate the quality of each algorithm.

However, we saw that PC and MIIC have parameters allowing to control the edge selection in order to make a more or less dense graph. Moreover, MIIC gives partial correlation and confidence score for each edge adding information to the relation, and PC gives pMax value. But the latter has a much higher execution time than MIIC. Even if HC made a dense graph and does not give such meaningful information about edges, it founds the main relations involving ***tp53*** and ***Ploidy***.

Finally, our github repository is open source and free to use[2]

References

[Sunil Kumar Raghavan Unnithan and Jathavedan, 2014] Sunil Kumar Raghavan Unnithan, B. K. and Jathavedan, M. (2014). Betweenness centrality in some classes of graphs. *International Journal of Combinatorics*, 2014:12.

[Verny et al., 2017] Verny, L., Sella, N., Affeldt, S., Singh, P., and Isambert, H. (2017). Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology*, 13:e1005662.

[2] https://github.com/MatPont/Network_Reconstruction_Analysis_TPs

Appendices

A MIIC graph with Cytoscape

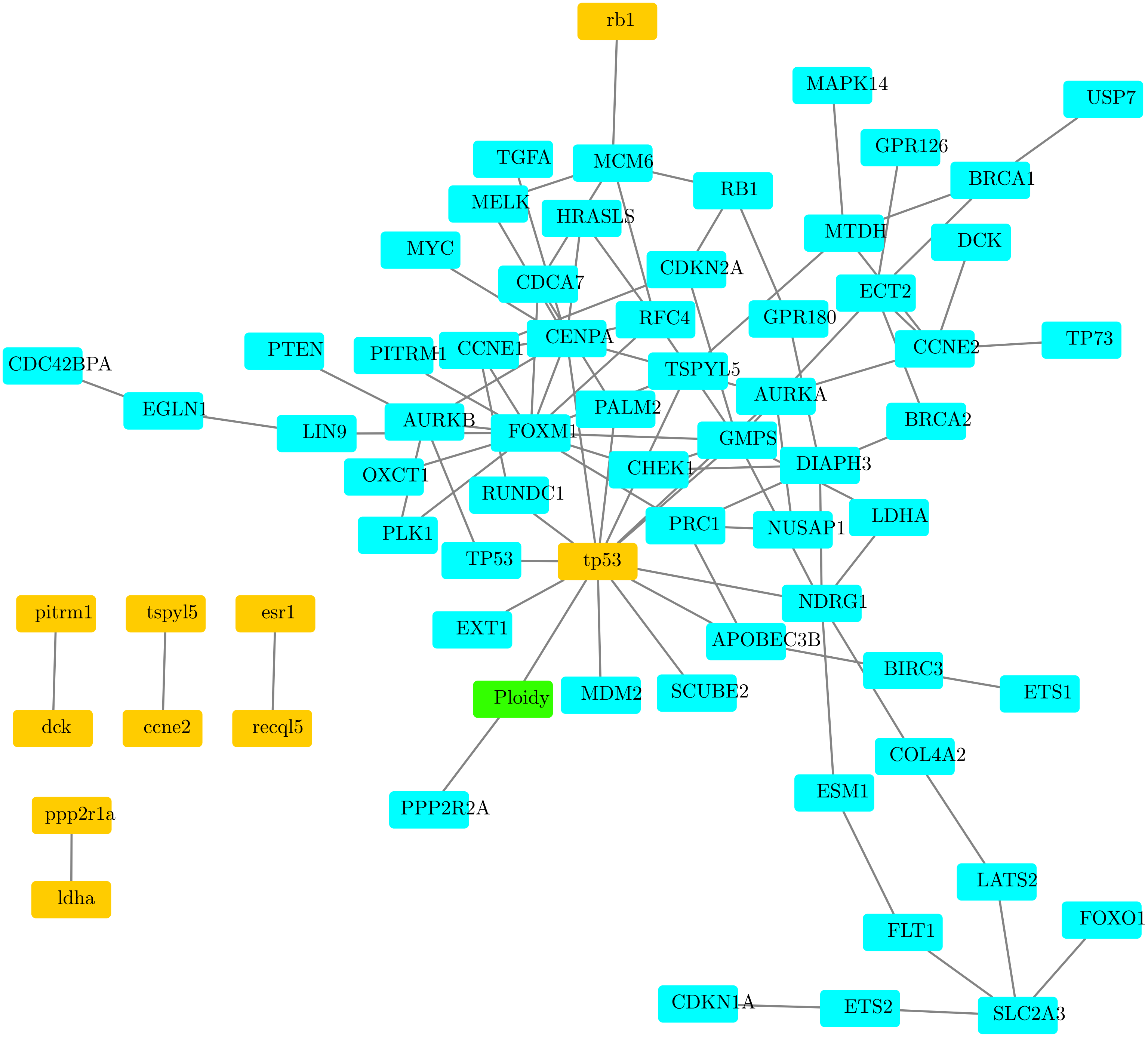


Figure 6: Reconstructed network using the MIIC algorithm with $confidenceShuffle = 100$ and $confidenceThreshold = 0.001$ represented with Cytoscape.